Published in final edited form as:

Methods Mol Biol. 2017 ; 1666: 441–453. doi:10.1007/978-1-4939-7274-6_21.

Calibrating Population Stratification in Association Analysis

Huaizhen Qin^{1,2} and Xiaofeng Zhu³

¹ Department of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, 70112, USA. hqin2@tulane.edu

² Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, 44106, USA. hqin2@tulane.edu

³ Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, 44106, USA

Abstract

In genetic association studies, it is necessary to correct for population structure to avoid inference bias. During the past decade, prevailing corrections often only involved adjustments of global ancestry differences between sampled individuals. Nevertheless, population structure may vary across local genomic regions due to the variability of local ancestries associated with natural selection, migration, or random genetic drift. Adjusting for global ancestry alone may be inadequate when local population structure is an important confounding factor. In contrast, adjusting for local ancestry can more effectively prevent false positives due to local population structure. To more accurately locate disease genes, we recommend adjusting for local ancestries by interrogating local structure. In practice, locus-specific ancestries are usually unknown and must be inferred. For recently admixed populations with known reference ancestral populations, locus-specific ancestries can be inferred accurately using some hidden Markov model-based methods. However, SNP-wise ancestries cannot be accurately inferred when ancestral population information is not available. For such scenarios, we propose employing local principal components (PCs) to present local ancestries and adjusting for local PCs when testing for gene–phenotype association.

Keywords

Genome-wide association studies; Migration; Random genetic drift; Natural selection; Admixed populations; Global ancestry; Local ancestries; Local principal components; Hidden Markov algorithms; Fine mapping

1 Introduction

In association studies, the concerns for population stratification can be dated back to 20 years ago [1]. It has been well realized that population stratification—systematic ancestry differences between study subjects—can confound association tests [2–13]. Population structure exists in study subjects as a result of distinct ethnic groups or a single pool of

admixed individuals. Global population structure can be characterized by individual global ancestries. An individual's global ancestry can be calculated as the proportions of his/her genome inherited from the underlying ancestral populations. Local population structure in a local genomic region or at a locus can be similarly characterized. During the past decade, statistical methods to control the false positive rate due to population stratification have mainly involved adjustments for global population structure, which is mainly due to recent migration and random genetic drift. Some prevailing paradigms are genomic control [3], structural association methods [4, 5] and principal component methods [6, 7, 10] that use markers randomly selected across the genome. In genomic control, the variance inflation factor of a test statistic is assumed to be constant across the entire genome. In a classical principal component (PC) method, the PCs of the genotype score matrix of genome-wide markers (referred to as global PCs) are used as ancestry surrogates in the association analysis for testing each marker. Many studies have shown that the global PCs can effectively represent human demographic history [6, 10–12, 14–16].

Nevertheless, subtle local structures do occur in some small genomic regions, owing to demographic history, natural selection pressure and random fluctuations of admixture [17, 18]. The imprint of natural selection, for example, has recently been identified in many regions across the genome [19, 20] and can create substantial variation in population differentiation, which in turn affects the degree of variance inflation for specific loci [21]. Even subtle structure, if ignored, can either inflate type I error or reduce statistical power, especially when the sample size is large [8].

For recently admixed populations with known reference ancestral populations (e.g., African-Americans), locus-specific ancestry can be inferred accurately using hidden Markov model-based methods [4, 22–27]. However, it is difficult to accurately infer locus-specific ancestries for an admixed individual owing to either the lack of ancestral population information or the high similarity between ancestral populations. For example, population admixture of European Americans occurred within populations of similar origin. For such scenarios, we propose using local PCs of the genotype score matrix of the markers within local genomic regions to represent local ancestries, at the same time using global PCs to represent global ancestries, and interrogating the ancestries in local genomic regions for fine mapping [13]. Extensive simulations and applications to the datasets from three genome-wide association studies illustrate the necessity and practical implications of the adjustment of local ancestry PCs. Both European Americans and African-Americans demonstrate greater variability in local ancestry than do Nigerians. Adjusting local PCs successfully eliminates the wellknown spurious association between the LCT gene and height in European Americans due to the underlying population structure. In this chapter, we illustrate how to run the local ancestry principal components correction (LAPCC) and hidden Markov algorithm ELAI [27].

In LAPCC, we divide each chromosome into adjacent 4 Mb segments (referred to as window cores hereafter) according to the SNP map of base pair positions. Typically, we add an envelope with an 8 Mb-margin to each side of a core (Fig. 1) and construct a 20 Mb-window. We choose this window width in accordance with the linkage disequilibrium due to recent population admixture [22, 24]. The left (right) envelopes of the first (last) two windows of some chromosomes might be shorter than 8 Mb. For each local window, we

compute the first $\ell = 10$ (see Note 1) PCs of the genotypic scores of window-wide SNPs to adjust for the population structure of the SNPs within the 4 Mb-core. At each SNP in the core, we compute genotype score residuals $\tilde{g}_i$ and trait value residuals $\tilde{y}_i$ by regressing genotype scores g_i and trait values y_i on the ten local PCs. We measure the evidence of phenotype-genotype association by $s^2 = (N - \ell - 2)r^2/(1 - r^2)$, where N is the number of individuals used after excluding individuals with missing genotypes and r is the correlation coefficient between the residuals $\tilde{y}_i$ and $\tilde{g}_i$. Asymptotically, s^2 follows a χ_1^2 distribution under the null of no genotype-phenotype association if all confounding factors are well adjusted.

To be specific, we denote by $\mathbf{G} = (g_{ij})$ the $M \times N$ genotypic matrix (of the entire genome or a local window), where $g_{ij} \in \{0, 1, 2\}$ is the copy number of a reference allele at the i th marker of the j th subject. We then center each row i of matrix $\mathbf{G}$ by the row mean $\mu_i = N^{-1} \sum_{j=1}^N g_{ij}$ and denote the centered matrix by $\mathbf{X} = (x_{ij})$. We exclude each missing entry g_{ij} from the computation of μ_i and set the corresponding x_{ij} to be 0. We denote the eigensystem of the $N \times N$ matrix $\mathbf{C} = \mathbf{X}'\mathbf{X}$ by $\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$, where $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_N]$, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N)$, and $\lambda_1 > \dots > \lambda_{N-1} > \lambda_N$ are eigenvalues corresponding to eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_N$, respectively, and in particular $\lambda_N = 0$ and $\mathbf{v}_N = N^{-1/2} (1, \dots, 1)'$. Following others' previous research work [10], we define the k th axis of variation as eigenvector $\mathbf{v}_k = (\mathbf{v}_{1k}, \dots, \mathbf{v}_{Nk})'$ and select κ eigenvectors (PCs) as the ancestral surrogates. We do not normalize the matrix $\mathbf{X}$ (see Note 2) and do not suggest thinning SNPs for calculating the PCs (see Note 3). Instead, we only center the genotypic scores and calculate the PCs from all available genotypic data.

To interrogate local structure, we calculate the coefficients of multiple-determination (R^2) and squared coefficients of canonical correlation (λ^2) between the global PCs and local PCs. Let N denote the sample size, $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$ denote the $N \times K$ matrix consisting of the first K global PCs, and $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$ denote the $N \times K$ matrix consisting of the first K local PCs in a local window. The coefficient of multiple-determination R_j^2 for $\mathbf{b}_j$ and $\mathbf{A}$ is the R^2 for the linear regression of $\mathbf{b}_j$ on $\mathbf{A}$. The j th largest squared coefficient of canonical correlation λ_j^2 between $\mathbf{A}$ and $\mathbf{B}$ is the j th largest coefficient of determination between any linear combination of $\mathbf{B}$'s columns and any linear combination of $\mathbf{A}$'s columns.

¹In LAPCC, setting $\ell = 10$ is somewhat arbitrary. Patterson et al. [32] developed a formal procedure to determine ℓ using the genotypic data. This procedure is based on the Tracy-Widom theory [32, 33] and proves to be conservative if the normalized genotypic data matrix is a true Wishart matrix [32]. Improvements may be possible since the first ℓ PCs may not be the most informative.

²According to our analysis of real data, normalization may hurt the representativeness of the global (local) PCs for the true ancestry when the data involve more than two ancestral populations (Fig. 3). For the case of only two ancestral populations, the results without normalization are mathematically valid. Thus, we give up normalization even if it may yield better results under certain conditions [32], which may be violated in practice.

³In our simulations, random correlations between markers have little impact on global PCs of tens of thousands to millions of genome-wide SNPs (Fig. 4). The first global PC coordinates of a large number of unlinked ancestral informative markers (AIMs) across the genome are highly correlated with true individual global ancestries. The first global PC coordinates using more random markers represent the true global ancestries even better, regardless of the abundant linkage disequilibrium (LD). This suggests that global PCs of all available genotypic data would well represent global ancestries. In the same vein, we do not suggest thinning SNPs to calculate the local PCs.

Mathematically, $R_1^2 \leq \lambda_1^2$, and $R_1^2 + \dots + R_K^2 = \lambda_1^2 + \dots + \lambda_K^2 \cdot R_j^2$ equals the j^{th} diagonal element and λ_j^2 equals the j^{th} largest eigenvalue of $\mathbf{B}'\mathbf{A}\mathbf{A}'\mathbf{B}$, and $R_j^2 = r_{j1}^2 + \dots + r_{jK}^2$, where $r_{ji}^2 = (\mathbf{b}_j'\mathbf{a}_i)^2$ is the coefficient of determination between $\mathbf{b}_j$ and $\mathbf{a}_i$. Canonical correlation analysis and multiple-determination analysis enable us to evaluate the degree of discrepancy between local and global PCs and facilitates the exploration of population structures. R^2 indicates how much variation of the local PCs can be accounted for by the global PCs and λ^2 measures the shared variance between the local and global PCs.

2 Methods

2.1 The Software Package LAPCC

We will illustrate how to use the software package LAPCC.exe with genome-wide genotypic data sets—the Maywood, Nigeria and Framingham data sets [28, 29]. The file LAPCC.exe can be downloaded from <http://darwin.cwru.edu/LAPCC/LAPCC.exe>. The user needs to put the file LAPCC.exe and the formatted data in the same folder

2.1.1 Formatting the Data—First we will explain how to preprocess and format the data. For example, the Maywood cohort comprises 775 unrelated AfricanAmericans from Maywood, IL, with 909,622 SNPs genotyped on the Affymetrix 6.0 platform. We drop 74 individuals because of possible DNA contamination, false identity, and relatedness. For the 701 retained individuals, we remove 86,800 SNPs with missing rates >5% or minor allele frequencies <1%. The final analysis data-set comprises 822,822 SNPs on each of 701 individuals. Similarly, after QC, the Nigeria Affymetrix 6.0 dataset contains 759,222 SNPs genotyped for 982 individuals. For the Framingham dataset, Mendelian errors are checked and the corresponding SNPs with Mendelian inconsistencies are set to missing. SNPs with departure from HWE P -values $<10^{-6}$ are dropped. We select unrelated individuals from each family (i.e., spouses) based on an algorithm that prioritizes individuals with higher genotyping rates, selecting individuals at random when needed. After QC, the Framingham dataset contains 415,281 SNPs genotyped for 1106 unrelated individuals

For each chromosome in each data set, five plain input files must be prepared like the following example. In the Maywood data set, the first autosome contains 67,242 SNPs. The first file is conf_chr1.txt, which contains three rows as shown below:

67242

711153

247165315

The first row is the number of SNPs on autosome 1, the second and third rows give the base pair positions of the first and the last SNPs on the autosome, respectively. The second input file is bp_chr1.txt, containing 67,242 rows. Each row contains the base pair position of one SNP and all the base pair positions are sorted in ascending order:

711153

730720
742429
751595
.....

The third file is snp_chr1.txt, containing the rs-numbers of all the 67,242 SNPs:

rs12565286
rs12082473
rs3094315
rs2286139
.....

The fourth file is geno_chr1.txt, containing a 701-by-67,242 white-spaced matrix of genotypic scores. For each SNP (column) a reference allele is randomly assigned and the copy number of the reference allele in each person is recorded as his/her genotypic score at the SNP. Missing genotypes are recorded as 9. In the file geno_chr1.txt, the genotypic scores of the first five SNPs of the first four individuals are as below:

1	2	2	2	2	...
2	1	2	2	2	...
2	1	0	1	2	...
2	-9	2	2	2	...
...	...	...	...	...	...

The last file is y_height.txt, listing the trait values (the residuals of height after adjusting for such covariates such as sex, age, and age-squared), one person per row:

0.710633
-0.346286
-1.25865
0.350496
.....

2.1.2 Run the package—After preparing the five plain files, the user can run LAPCC.exe to process all or specified autosomes in a parallel way. For each specified autosome, the LAPCC.exe outputs windows, window-wide R^2 - and λ^2 -values as well P

values of all input SNPs on the autosome. For the three aforementioned populations, we compare local and global PCs to uncover local ancestry patterns. Figure 2 presents the λ^2 -values between window-wide local PCs and the first ten global PCs. In general, local PCs and global PCs here only display a small degree of correlation. However, the Maywood and Framingham datasets demonstrate substantially larger λ^2 -values and larger variation in λ^2 -values than does the Nigerian dataset. These results suggest that there is relatively little population structure in the Nigerian sample, whereas Maywood and Framingham display relatively more population structure. Somewhat surprisingly, the Framingham dataset demonstrates much more complex local population structure than do the other two datasets.

2.2 The ELAI Software

In LAPCC, the window-wide PCs are easy to compute. However, such local PCs would be too rough to accurately estimate the local ancestries when sample size and window size are small (see Notes 4 and 5). For each admixed individual, SNP-wise ancestries can be inferred from the genotype data by the ELAI software [27]. For illustration purposes, here we detail the steps for inferring SNPwise local ancestries of the African-Americans in the Maywood cohort study.

2.2.1 Download the Phased Reference Panels—Typically, CEU (Centre d'Etude du Polymorphisme Humain in Utah, USA) and YRI (Yoruban in Ibadan, Nigeria) are taken as the two ancestral populations of African-Americans. The phased data of the hapmap3 reference panels can be downloaded from the link: ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3_r3/plink_format/. Only unrelated individuals in the panels will be used as references (112 subjects for CEU, and 147 for YRI). The second generation PLINK [30] is employed to extract the reference data on the unrelated individuals:

```
./plink--filehapmap3_r3_b36_fwd.CEU.qc.poly--keepCEU_Unrelated.
txt--make-bed--outCEU_all-noweb
./plink--file./plink--filehapmap3_r3_b36_fwd.YRI.qc.poly--keep
YRI_Unrelated.txt--make-bed--outYRI_all-noweb
```

The plain text files CEU_Unrelated.txt and YRI_Unrelated.txt contain the information of the unrelated individuals and their family IDs.

2.2.2 Format the unrelated reference panels—Local ancestry at each SNP for all the individuals can be inferred for each chromosome separately. For illustration purposes only, we take chromosome 22 as an example.

⁴The local windows are often empirically created and thus maybe away from the optimal. Standard local PCs may be fragile to apparent random inter-SNP correlations, especially for local genomic regions without sufficient ancestral informative markers. Thus, for local PCs, one may weight genotypic scores of the SNPs within a local window [34]. SNP-specific ancestries can be inferred using ELAI [27] if information of ancestral populations is known and reference panels are available. It is reasonable to incorporate the adjustment of global PCs to prevent false positives due to other sources of confounding.

⁵LAPCC was designed for datasets comprising unrelated individuals and would not account for the other types of sample structure, i.e., cryptic relatedness and family structure. Cryptic relatedness may occur in a wide range of datasets and it is necessary to model family structure in family-based association studies with sample ascertainment [35]. It may be instructive to incorporate the full covariance structure across individuals using mixed models.

1. *Extract chromosome-wise reference data.* The chromosome-wise reference data can be extracted by running the following command lines:

```
./plink--bfileYRI_all--chr22--outYRIChr22--make-bed-noweb
./plink --bfile YRIChr22 --recode --out YRIChr22_new -noweb
./plink--bfileCEU_all--chr22--outCEUChr22--make-bed-noweb.
./plink --bfile CEUChr22 --recode --out CEUChr22_new -noweb
```

Six files will be generated: YRIChr22_new.ped, YRIChr22_new.map, YRIChr22_new.log; CEUChr22_new.ped, CEUChr22_new.map, CEUChr22_new.log.

2. *Convert the reference PED data to BIMBAM files.* The fcGENE software [31] is employed for the data conversion:

```
./fcgene--fileCEUChr22_new--oformatbimbam--outCEUChr22_Ref
./fcgene--fileYRIChr22_new--oformatbimbam--outYRIChr22_Ref
```

Running the above command lines will generate CEUChr22_Ref.geno.txt, YRIChr22_Ref.geno.txt and two SNP-position files. For CEUChr22_Ref.geno.txt, we replace the first row “112” by “112=”, and replace the second row “0” by the number of SNPs on chromosome 22, which is “19,711”. The equality after the sample size “112” indicates that the reference data are phased. The YRIChr22_Ref.geno. txt file needs to be modified similarly. The Linux command lines are:

```
sed -i '1 s/112/112 =/' CEUChr22_Ref.geno.txt
sed -i '/^0/s/0/19711/' CEUChr22_Ref.geno.txt cate
sed -i '1 s/147/147 =/' YRIChr22_Ref.geno.txt
sed -i '/^0/s/0/19711/' YRIChr22_Ref.geno.txt
```

2.2.3 Format the Study Data—Similarly, the data of chromosome 22 are extracted from the Maywood dataset and converted to BIMBAM files. The following command lines are employed to extract chromosome-wise study data:

```
./plink--bfileData_701_qc--chr22--outWood701Chr22--make-bed -noweb
./plink--bfileWood701Chr22--recode--outWood701Chr22_new-noweb
```

Next the fcgene software [31] is employed to convert the study PED data Wood701Chr22_new to BIMBAM files:

```
./fcgene --file Wood701Chr22_new --oformat bimbam --out Wood701_Chr22
```


Running the above command lines will generate Wood701_Chr22.geno.txt and Wood701_Position_Chr22.txt. Since the study data are raw genotype data, only the second row “0” of Wood701_Chr22.geno.txt needs to be replaced with “11,081”, the number of SNPs on Wood701Chr22_new. The following Linux command line applies:

```
sed -i '/^0/s/0/10811/' Wood701_Chr22.geno.txt
```

The position file Wood701_Position_Chr22.txt needs to be modified as below. The first column of the file is the SNP ID (rs), the second column is its physical location, and the third column is its chromosome number. For a SNP, the rs number, physical position and chromosome ID must be comma delimited.

2.2.4 Infer SNP-wise ancestry scores by ELAI—After preparing all required files, running the following ELAI command line:

```
./elai-lin -g CEUChr22_Ref.geno.txt -p 10 -g YRIChr22_Ref.geno.txt -p 11 -g  
Wood701_Chr22.geno.txt -p 1 -pos Wood701_Position_Chr22.txt -s 30 -C 2 -c 10  
-o AdmWood701_Chr22 -mixgen 50 --exclude-nopos --exclude-missl
```

One of the output files, AdmWood701_Chr22.ps21.txt, contains the local ancestry scores of all the admixed individuals. Each row is for one individual and every two columns are for one SNP. The odd columns are the CEU local ancestry scores.

Acknowledgments

This work was funded in part by NHGRI grant HG003054 to X.Z. and by Tulane’s Committee on Research fellowship (600890) and Carol Lavin Bernick Faculty Grant (632119) to H.Q.

References

1. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 43:520–526 [PubMed: 3177389]
2. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* 265:2037–2048 [PubMed: 8091226]
3. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004 [PubMed: 11315092]
4. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67:170–181 [PubMed: 10827107]
5. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 68:466–477 [PubMed: 11170894]
6. Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. *Genet Epidemiol* 23:181–196 [PubMed: 12214310]
7. Zhang S, Zhu X, Zhao H (2003) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol* 24:44–56 [PubMed: 12508255]

8. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nat Genet* 36:512–517 [PubMed: 15052271]
9. Campbell CD et al. (2005) Demonstrating stratification in a European American population. *Nat Genet* 37:868–872 [PubMed: 16041375]
10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909 [PubMed: 16862161]
11. Zhu X et al. (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* 82:352–365 [PubMed: 18252216]
12. Zhu X et al. (2008) Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet* 60:547–569 [PubMed: 18358332]
13. Qin et al. (2010) Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* 26 (23):2961–2968 [PubMed: 20889494]
14. Cavalli-Sforza LL, Bodmer WF (1999) *The genetics of human populations*. Dover, Mineola, New York
15. Epstein MP et al. (2007) A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 80:921–930 [PubMed: 17436246]
16. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 40:646–649 [PubMed: 18425127]
17. Tang H et al. (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet* 81(3):626–633 [PubMed: 17701908]
18. Genovese G et al. (2010) Association of trypanolytic ApoL1 variants with kidney disease in African-Americans. *Science* 7:1–7
19. Voight BF et al. (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72 [PubMed: 16494531]
20. Sabeti PC et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918 [PubMed: 17943131]
21. Crow JF, Kimura M (1970) *An introduction to population genetics theory*. Harper & Row, New York, pp 469–478
22. Patterson N et al. (2004) Methods for high density admixture mapping of disease genes. *Am J Hum Genet* 74:979–1000 [PubMed: 15088269]
23. Tang H et al. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 79:1–12 [PubMed: 16773560]
24. Zhu X et al. (2006) A classical likelihood based approach for admixture mapping using EM algorithm. *Hum Genet* 120:431–445 [PubMed: 16896924]
25. Sankararaman S et al. (2008) Estimating local ancestry in admixed populations. *Am J Hum Genet* 82:290–303 [PubMed: 18252211]
26. Price AL et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5:e1000519 [PubMed: 19543370]
27. Guan Y (2014) Detecting Structure of Haplotypes and Local Ancestry. *Genetics* 196 (3):625–642 [PubMed: 24388880]
28. Kang SJ et al. (2010) Genome wide association of anthropometric traits in African and African derived populations. *Hum Mol Genet* 19 (13):2725–2738 [PubMed: 20400458]
29. Levy D et al. (2009) Genome-wide association study of blood pressure and hypertension. *Nat Genet* 41:677–687 [PubMed: 19430479]
30. Chang CC, Chow CC, Tellier L, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4(7):1–16 [PubMed: 25838885]
31. Roshyara NR, M Scholz M (2014) AfcGENE: a versatile tool for processing and transforming SNP datasets. *PLoS One* 9(7):e9758
32. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):2074–2093., e190. doi:10.1371/journal.pgen.0020190
33. Johnstone I (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* 29:295–327

34. Zou F et al. (2010) Quantification of population structure using correlated SNPs by shrinkage principal components. *Hum Hered* 70:9–22 [PubMed: 20413978]
35. Price AL et al. (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews* 11:459–463

Fig. 1.
A typical window consists of a 4 Mb-core and an envelope with an 8 Mb-margin to each side of the core. The first ℓ PCs of the genotypic score matrix of the SNPs in the 20 Mb-window are employed to calibrate the local ancestries of the SNPs within the 4 Mb-core

Fig. 2.

The distributions of the λ^2 -values of the local windows in three GWAS datasets. For each window in a given genotype dataset, λ^2 is the largest squared coefficient of canonical correlation between the first ten window-wise PCs and the first ten global PCs. The Maywood participants demonstrate apparent population structure, the Nigerian samples demonstrate little population structure, whereas the Framingham participants demonstrate much more complex local population structure than do the other two samples

Fig. 3. PCs with (a) and without (b) normalizing the GAW17 genotypic data of 697 unrelated individuals. Clearly, the PCs without normalization provide better discrimination, although both analyses roughly classify the 697 individuals into three big groups: CEPI and Tuscan; Luhya and Yoruba; and Denver Chinese, Han Chinese, and Japanese. The PCs without normalization appears more robust to outliers

Fig. 4.

Pearson coefficients of correlation between the standard global PC1 coordinates of distinct subsets of 1,969,739 SNPs and individual global ancestries of the 2000 individuals. The data set is generated by the GenoAnceBase0 program [13] to the CEU and YRI haplotypes of the HapMap data (Phase II) to simulate African-American genomes. The standard global PC1 coordinates of the 3029 unlinked AIMs across the genome are highly correlated to true individual global ancestries. The standard global PC1 coordinates using more random markers represent the true global ancestries even better regardless of even more abundant LD